Plataforma de Datos Abiertos Enlazados para la Gestión y Visualización de Datos Primarios de Ciencias del Mar

Carlos Buckle[1,2], Marcos Zarate[1,2,4], Renato Mazzanti[1,2,3], Claudio Delrieux[2,5], Mirtha Lewis[4]

[1] Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB) Sede Puerto Madryn, Chubut, Argentina,+54 280-4883585 Int. 117.

[2] Laboratorio de Investigación en Informática de la UNPSJB (LINVI-UNPSJB)

[3] Unidad de Gestión de la Información, Centro Nacional Patagónico, Consejo Nacional de Investigaciones Científicas y Técnicas (UGI-CENPAT-CONICET)

[4] Centro para el Estudio de Sistemas Marinos, Centro Nacional Patagónico, Consejo Nacional de Investigaciones Científicas y Técnicas (CESIMAR-CENPAT-CONICET)

[5] Depto. de Ingeniería Eléctrica y de Computadoras, Universidad Nacional del Sur (DIEC-UNS)

cbuckle@unpata.edu.ar, zarate@cenpat-conicet.gob.ar, renato@cenpat-conicet.gob.ar

Resumen

La gestión, preservación y disposición con acceso público de datos primarios producidos en investigaciones están cobrando cada vez mayor importancia. Nuestro país, define normativas de ciencia abierta enfocadas al uso democrático y eficiente del conocimiento científico, para potenciar el desarrollo y como herramienta para un mejor aprovechamiento de los recursos utilizados en la exploración y recopilación de datos científicos. Una necesidad común es integrar información de diferentes repositorios y unificarla bajo un mismo vocabulario con conceptos relacionados, para ser explorada por grupos de investigación, citada en producciones científicas, visualizada para su interpretación y sistematizada bajo estándares que permitan la interoperabilidad con otras aplicaciones o repositorios.

Este proyecto propone investigar posibles soluciones a estas necesidades aplicándolas al dominio de las ciencias del mar. Para ello, se dispone de información recolectada en campañas multidisciplinarias sobre el Atlántico Sur Occidental. Como aplicación se propone el desarrollo de una plataforma para la especificación, modelado, generación y publicación de los conjuntos de datos primarios como datos abiertos enlazados, con visualizaciones que faciliten su interpretación y comparación, como así también su explotación basada en semántica.

Palabras clave: Datos Primarios Científicos, Datos Enlazados, Visualización de Datos, Repositorios Digitales.

Contexto

El proyecto propone avanzar sobre resultados y líneas de investigación del proyecto precedente *UNPSJB-PI 1424-Infraestructura de Acceso a Datos Primarios con aporte de semántica en Repositorios Digitales*, en el cual se identificaron ventajas en los Datos Abiertos Enlazados (Linked Open Data) [1] como camino estándar hacia la integración de datos abiertos [2] heterogéneos de diferentes dominios, bajo vocabularios comunes y con posibilidades de razonamiento por partes de agentes de software. Este trabajo se focalizará específicamente en la

explotación de datos relacionados con las ciencias del mar y se desarrollarán visualizaciones adaptadas a las necesidades de los investigadores.

Esta propuesta tiene una concepción interdisciplinaria y por ello el grupo de trabajo incluye: investigadores de las ciencias de la computación de diferentes áreas (web semántica, bases de datos, visualización científica e inteligencia artificial) e investigadores de las ciencias biológicas, expertos en el dominio de aplicación.

El proyecto se inscribe dentro del Laboratorio de Investigaciones en Informática (LINVI-UNPSJB) y se integra con otros grupos de investigación de la Universidad Nacional del Sur (UNS), del Centro Nacional Patagónico (CENPAT-CONICET) y del CIT Golfo San Jorge (UNPSJB-UNPA-CONICET). Está avalado por el Consejo Directivo de la Facultad de Ingeniería de la UNPSJB y será financiado por la Secretaria de Ciencia y Técnica de la UNPSJB para llevar a cabo durante 2020-2022.

1. Introducción

La complejidad y el alcance de la recolección de datos de campañas oceanográficas exigen una aproximación interdisciplinaria y una proyección amplia en el uso de la información obtenida. Estas campañas utilizan el Sistemas Nacionales de Datos del Mar (SNDM) [3] como repositorio final de datos, y requieren de una gestión integrada, tanto para su tratamiento como para su explotación, de manera que sea posible generar productos de síntesis que faciliten la tarea de análisis y descubrimiento de conocimiento.

SNDM utiliza como Plataforma el software provisto por *International Oceanographic Data and Information Exchange* (IODE). Su implementación en Argentina no fue del todo exitosa y tiene algunas limitaciones:

- Brinda la información de los diferentes recursos en forma fragmentada y no hay recursos suficientemente detallados para el ingreso de datos biológicos.
- No disponen de un desarrollo para el manejo restringido de los datos, para que durante el periodo de embargo post campaña, los datos puedan ser compartidos entre los grupos de trabajo o proyectos en cooperación.
- No existe la posibilidad de generar un informe con los datos más relevantes de una campaña de investigación en el mar.
- No permite interpretar la información de manera visual, por ejemplo un mapa interactivo donde el usuario pueda realizar búsquedas personalizadas.

La ciencia de datos es considerada una disciplina fundamental para abordar la complejidad y el alcance de las temáticas que exigen una aproximación interdisciplinaria y una proyección amplia en el uso de la información. En el Atlántico Sur las mediciones de este tipo son escasas y no se dispone de un sistema de captura de información adecuado. Por lo tanto, es necesario desarrollar sistemas capaces de gestionar su integración y su comunicación, tanto para un aprovechamiento integral y secundario de los grupos e instituciones participantes como para usuarios externos que requieran de información.

La relevancia de la propuesta se identifica en diferentes niveles: a nivel tecnológico, la creación de una plataforma de Datos primarios Abiertos Enlazados permitirá resolver los problemas de integración de datos heterogéneos bajo un mismo vocabulario [4] y disponerlos para ser explotados mediante búsquedas semánticas [5] y consultas basadas en lenguaje natural. Complementariamente, el desarrollo de Visualizaciones de Datos del Mar (Data Visualización) [6] facilitará su interpretación y comparación.

A nivel de política científica para la región y el país, se logrará un prototipo de servicio estandarizado para repositorios de datos biológicos del Atlántico Sur Occidental, con información unificada provenientes de

diferentes plataformas de muestreo (buques oceanográficos, ROVs, submarinos tripulados, etc.), que podrá ser reusada, citada, explotada para la generación de nuevos conocimientos.

A nivel socio-económico, las posibilidades de reuso de datos permitirán reducir los altos costos de las campañas de investigación en el mar y aportará al uso democrático y eficiente de datos científicos.

A nivel de formación de recursos humanos del proyecto participan estudiantes de posgrado de la disciplina Ciencias de la Computación, investigadores en formación y estudiantes avanzados de la Licenciatura en Informática de la UNPSJB.

2. Motivación

Dada la importancia del acceso abierto a datos científicos recogidos en el Atlántico Sur Occidental, tanto para la investigación y el desarrollo (I+D) en el país, como para garantizar visibilidad en el contexto regional e internacional; y dado que los datos relacionados mediante Datos Abiertos Enlazados pueden proporcionar la forma de vincular campañas del mar con publicaciones científicas, conjuntos de datos biológicos y químicos a través de vocabularios controlados y ontologías, esta investigación considera relevante plantear como hipótesis que la plataforma de Datos Primarios Enlazados posibilita la realización de un modelo conceptual para la construcción de una plataforma web (prototipo) capaz de publicar y vincular las campaña sobre el mar argentino para que puedan ser compartidas y reutilizadas a nivel nacional e internacional. Esto conlleva a la necesidad de estudiar infraestructuras para la publicación de datos científicos [7] y a considerar que la aplicación de visualizaciones de datos colaborará con la tarea analítica, mediante herramientas gráficas que faciliten la tarea de interpretación y contrastación de datos provenientes de los repositorios.

3. Líneas de Investigación y Desarrollo

Este proyecto desarrolla como principal línea de investigación, el Modelado Conceptual en la Web Semántica [8] y la construcción de grafos de conocimiento oceanográfico [9, 10] mediante datos enlazados [11, 12] para la integración de datos científicos y su explotación [13, 14]. Pero además, será necesario abordar el tratamiento de grandes volúmenes de datos oceanográficos y meteorológicos [15], caracterizados por las 5V (Volumen, Velocidad, Variedad, Verosimilitud y Valor) de Big-data [16] y teorías, técnicas y herramientas para la visualización de datos científicos oceanográficos [17, 18].

4. Resultados esperados

El propósito de la investigación es desarrollar una plataforma de Datos Abiertos Enlazados para la gestión, explotación, vinculación y visualización de datos del mar recolectados en campañas sobre el Atlántico Sur Occidental.

Se busca proporcionar una gestión confiable de datos para garantizar la preservación y el acceso a nuestros activos nacionales de investigaciones sobre el mar Argentino. Con objetivos de democratización del conocimiento, incentivo a los grupos de investigación para sumar contribuciones y optimización de costos basada en reuso de datos. Se esperan como resultados:

1. Estudio del estado del arte respecto de las bases de datos del mar en general y en particular las que registran datos del mar Argentino, con posterior registro de metadatos a nivel de crucero oceanográfico (trayectorias, instrumentos utilizados, origen, destino, estaciones, etc.).

2. Especificación, Modelado, Generación y Publicación en una Plataforma de Datos Abiertos Enlazados, escalable, estándar y con incorporación de semántica del dominio de Ciencias del Mar.

3. Desarrollo de visualizaciones de datos (scientific data visualization) para la interpretación, contrastación y comparación, incorporando consultas en lenguaje natural.

4. Formación y Transferencia: Consolidación del grupo de investigación que trabaja en repositorios de datos primarios con semántica. Formación de estudiantes, doctorandos y jóvenes investigadores. Difusión de resultados y publicaciones. Transferencia a iniciativas similares.

5. Formación de recursos humanos

En este proyecto participa un docente de UNPSJB-Puerto Madryn, doctor en Ciencias de la Computación y actual becario posdoctoral CONICET focalizado en el desarrollo de Datos Enlazados para la gestión integrada de datos científicos multidisciplinares de ciencias oceanográficas, de biodiversidad y ambientales. También participa un becario doctoral CONICET perteneciente al CIT Golfo San Jorge que estudia y propone modelos de visualización de datos científicos. Sus directores son investigadores de la Universidad Nacional del Sur y del CESIMAR-CENPAT-CONICET, quienes también integran el equipo de trabajo en calidad de asesores y expertos de dominio.

Además, el equipo de trabajo incluye a seis docentes del Departamento de Informática de la Facultad de Ingeniería de la UNPSJB-Puerto Madryn. Uno de ellos iniciando una especialización en Gestión de Información Científica y Tecnológica y un joven investigador que estudiará explotación de grafos de conocimiento mediante visualizaciones.

También forman parte del proyecto 2 (dos) alumnos del ciclo superior de la carrera Licenciatura en Informática que podrán desarrollar su trabajo de tesina y realizar Instancias Supervisadas de Formación en la Práctica Profesional en el marco de este proyecto.

Referencias

[1] Tom Heath and Christian Bizer. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136, 2011.

[2] Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. Data sharing by scientists: Practices and perceptions. *PLoS ONE*, 6, 06/2011 2011.

[3] Sistema Nacional de Datos del Mar, Ministerio de Ciencia Tecnología e Innovación Productiva Presidencia de la Nación Argentina. *http://www.datosdelmar.mincyt.gob.ar*. Accedido: 15/2/2020.

[4] CL Chandler, RC Groman, A Shepherd, MD Allison, D Kinkade, S Rauch, PH Wiebe, and DM Glover. Using controlled vocabularies and semantics to improve ocean data discovery. In *AGU Fall Meeting Abstracts*, 2013.

[5] Judith A Blake and Carol J Bult. Beyond the data deluge: data integration and bioontologies. *Journal of biomedical informatics*, 39(3):314–320, 2006.

[6] Reiner Schlitzer. Interactive analysis and visualization of geoscience data with ocean data view. *Computers & geosciences*, 28(10):1211–1218, 2002.

[7] Giuseppe Andronico, Valeria Ardizzone, Roberto Barbera, Bruce Becker, Riccardo Bruno, Antonio Calanducci, Diego Carvalho, Leandro Ciuffo, Marco Fargetta, Emidio Giorgio, Giuseppe Rocca, Alberto Masoni, Marco Paganoni, Federico Ruggieri, and Diego Scardaci. e-infrastructures for e-science: A global view. *Journal of Grid Computing*, 9:155– 184, 06 2011.

[8] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284(5):34–43, May 2001.

[9] Marcos Zárate, Pablo Rosales, Pablo Fillottrani, Claudio Delrieux, and Mirtha Lewis. Oceanographic data management: Towards the publishing of pampa azul oceanographic campaigns as linked data. *In Proceedings of the 12th Alberto Mendelzon International Workshop on Foundations of Data Management (AMW 2018)*, 2018.

[10] Marcos Zárate, Pablo Rosales, Germán Braun, Mirtha Lewis, Pablo Rubén Fillottrani, and Claudio Delrieux. Oceangraph: Some initial steps toward a oceanographic knowledge graph. *In Iberoamerican Knowledge Graphs and Semantic Web Conference*, pages 33–40. Springer, 2019.

[11] Krzysztof Janowicz, Pascal Hitzler, Benjamin Adams, Dave Kolas, II Vardeman, et al. Five stars of linked data vocabulary use. *Semantic Web*, 5(3):173–176, 2014.

[12] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI Global, 2011.

[13] Tony Hey, Stewart Tansley, and Kristin Tolle. *The Fourth Paradigm: DataIntensive Scientific Discovery*. Microsoft Research, October 2009.

[14] Adam Leadbetter, Robert Arko, Cynthia Chandler, Adam Shepherd, and Roy Lowry. Linked Data An Oceanographic Perspective. *The Journal of ocean Technology*, 8(3), 2013.

[15] Tanu Malik and Ian Foster. Addressing data access needs of the long-tail distribution of geoscientists. In *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, pages 5348–5351. IEEE, 2012.

[16] Mark A Beyer and Douglas Laney. The importance of 'big data': a definition. *Stamford, CT: Gartner*, pages 2014–2018, 2012.

[17] Peter Fox and James Hendler. Changing the equation on scientific data visualization. *Science*, 331(6018):705–708, 2011.

[18] Nikos Bikakis and Timos Sellis. Exploration and visualization in the web of big linked data: A survey of the state of the art. *arXiv preprint arXiv*:1601.08059, 2016.